Latent space interpretation and visualisation for understanding the decisions of convolutional variational autoencoders trained with EEG topographic maps[⋆]

Taufique Ahmed[1,2,*,†], Luca Longo[1,2,†]

[1] Artificial Intelligence and Cognitive Load Research Lab

[2] School of Computer Science, Technological University Dublin

Abstract

Learning essential features and forming simple representations of electroencephalography (EEG) signals are difficult problems. Variational autoencoders (VAEs) can be used with EEG signals to learn the salient features of EEG data. But explainability should disclose knowledge of how the model makes its decision. The key contribution of this research is the combining of known components in a pipeline that allows us to give meaningful visualisations that help us understand which component of latent space is responsible for capturing which region of brain activation in EEG topographic maps. The results reveal that each component in the latent space contributes to capturing at least two generating factors in topographic maps. This pipeline can be used to produce EEG topographic maps of any scale. Furthermore, assist us in understanding each component of latent space responsible for activating a portion of the brain.

Keywords

Electroencephalography, Convolutional variational autoencoder, latent space interpretation, deep learning, spectral topographic maps

1. Introduction

Electroencephalography (EEG) is a method of recording brain activity (electrical potentials) using electrodes placed on the scalp [1]. Some research, for example, has transformed EEG signals into topographic power head maps to preserve spatial information [2]. Convolutional neural networks are frequently employed to reduce their dimensionality and automatically learn essential features [3]. An Autoencoder (AE) is a deep learning neural network architecture that uses unsupervised learning to learn efficient codings without the usage of labelled input [4]. A Variational Autoencoder (VAE) is a form of autoencoder that creates a probabilistic model of the input sample and then reconstructs it using that model. VAEs have shown a wide application

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

[*] Corresponding author.

[†] These authors contributed equally.

taufique.ahmed@tudublin.ie (T. Ahmed); luca.longo@tudublin.ie (L. Longo)
lucalongo.eu/about (L. Longo)
0000-0001-6332-7594 (T. Ahmed); 0000-0002-2718-5426 (L. Longo)

CEUR Workshop Proceedings (CEUR-WS.org)

with electroencephalographic (EEG) signals [5, 6, 7]. However, research into interpreting the latent space of a variational autoencoder to determine the importance of each latent space component in capturing the generating factor in spatially preserving EEG topographic maps is limited. In this study, the goal is to tackle the research problem to learn the importance of each latent component of VAE, trained with spectral topographic EEG maps for capturing the generative factors of EEG data. Therefore, the research question being addressed is: *Can we understand the reconstruction capacity of a convolutional variational autoencoder trained with spectral topographic maps by interpreting and visualising its learnt latent space representation?* The rest of the work is organised as follows. Section 2 investigates related work, whereas Section 3 describes an empirical study and its methodology. Section 4 presents the experimental results and findings. Finally, Section 5 concludes the manuscript by describing the contribution to the body of knowledge and highlighting future work directions.

2. Related Work

Traditional Autoencoders (AE) aim to learn prominent latent representations from unlabeled input while ignoring irrelevant features. Variational Autoencoders (VAEs) was recently proposed as an effective extension of AEs, for modeling a data's probability distribution and learning a latent space, usually of a lower dimension. It is ideal for unsupervised learning to understand the impact and importance of each latent component for capturing the number of true generative factors. VAE-based latent space analysis and decoding of EEG signals are important since they can precisely define and determine the latent relevant features [8]. VAE has been constructed with two distinct encoders to map the input into Zs and Zu, respectively, and then deliver the concatenated code to the decoder to reconstruct the input[9]. Another study used VAE and manually adjusted the latent activations, allowing the user to see the effect of different latent values on the generated output [10]. After all, if we want learned latent space representation to be interpretable, the latent component must have clear-cut meaning [11]. The researcher also illustrated how a VAE model's latent space may be made more explainable by utilizing latent space regularisation [12, 13, 14]. The majority of time series data are mapped to prominent representational characteristics and interpretation of its latent space results in improved clustering performance [15]. Because the disentanglement of latent space performs the clustering operation, no further clustering approach is required [16, 17]. Despite broad application and study into the interpretation of latent space, knowing VAE reconstruction decisions and the impact of its components on capturing the number of genuine generating elements remains inadequate.

3. Research design and methodology

In this study, if CNN-VAE is trained with spatially preserved EEG topographic maps and its interpretation of the learned latent space representation provides the knowledge of how well each component of the latent representation contributes to capturing the number of true generative factors in spatially preserving EEG topographic maps via visual plausibility. The detailed design of this research is illustrated in figure 1.

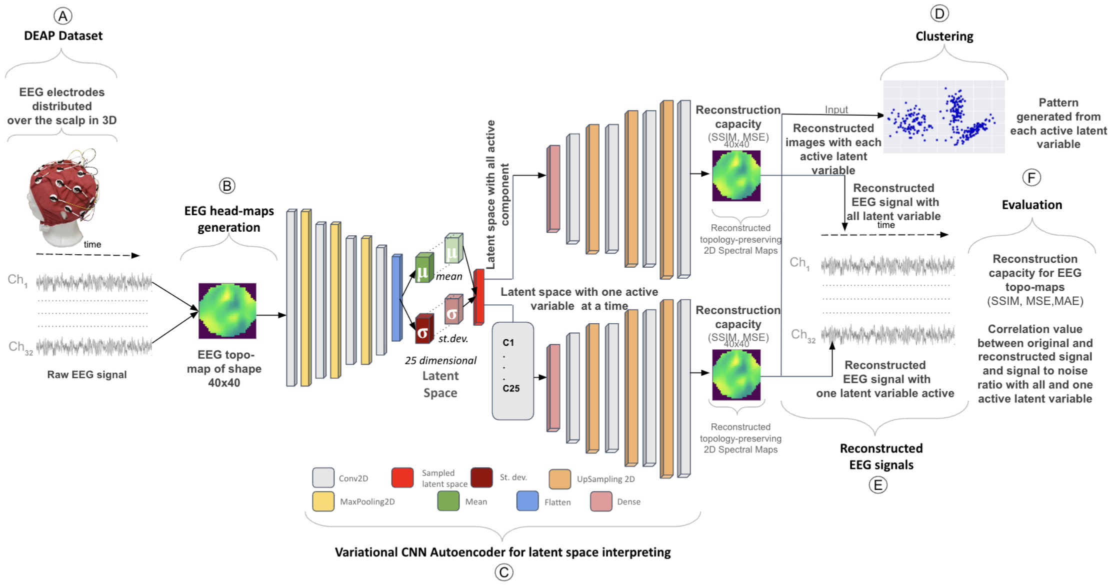
A
DEAP Dataset
EEG electrodes distributed over the scalp in 3D
time
Ch1
Ch32
Raw EEG signal
B
EEG head-maps generation
EEG topo-map of shape 40x40
μ
mean
σ
st.dev.
25 dimensional Latent Space
Latent space with all active component
Latent space with one active variable at a time
C1
C25
Reconstruction capacity (SSIM, MSE)
40x40
Reconstructed topology-preserving 2D Spectral Maps
Input
Reconstructed images with each active latent variable
D
Clustering
Pattern generated from each active latent variable
Reconstructed EEG signal with all latent variable
Reconstructed EEG signal with one latent variable active
Reconstructed EEG signals
E
F
Evaluation
Reconstruction capacity for EEG topo-maps (SSIM, MSE,MAE)
Correlation value between original and reconstructed signal and signal to noise ratio with all and one active latent variable
Conv2D
Sampled latent space
St. dev.
UpSampling 2D
MaxPooling2D
Mean
Flatten
Dense
Variational CNN Autoencoder for latent space interpreting
C

Figure 1: A pipeline for spatially-preserving EEG topographic maps generation and interpreting the latent space of CNN-VAE via visual plausibility.

The DEAP dataset was chosen because it contains multi-channel EEG recordings with 32 participants who watched 40 one-minute music video clips and tasks [18]. These EEG signals were transformed into 40×40 interpolated topographic maps that preserve spatial information about brain activation [19], a similar experiment was carried out in another investigation [20], as illustrated in (figure 1, B). Following the creation of the topographic maps, a Convolutional Variational Autoencoder (CNN-VAE) is built. The encoder network of CNN-VAE takes a 40×40 tensor (as seen in figure 1, C) and defines the approximate posterior distribution $Q(Z \mid x)$. The CNN-VAE decoder is a generative network that takes a latent space Z as input and returns the reconstructed EEG topographic maps. The architecture (figure 1, C) is made up of four 2D convolutional layers, each followed by a max pooling layer to minimize the dimension of the feature maps. In each convolutional layer, ReLU is employed as the activation function. To avoid overfitting, an early stopping strategy with a patience value of ten epochs is used, which indicates that training is stopped if the validation loss does not improve for ten consecutive epochs. To examine the number of generative factors captured from each active latent component, the reconstructed EEG topographic map from the decoder of CNN-VAE with latent space representation of only one active component is passed as an input to the K-means algorithm. In terms of how well samples are clustered with other samples that are similar to each other, the silhouette score is used to evaluate the quality of clusters generated using clustering methods such as K-Means. The reconstruction capacity of this model is evaluated by Structural Similarity Index (SSIM), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) derived for the reconstructed topographic maps.

4. Results & Discussion

The reconstruction capacity of the CNN-VAE model and its explainability is described with two different scenarios because it contains two decoder networks. One trained with all latent components, whereas the other trained with only one active latent at a time. In the second scenario, interpreting the disentangled representation of CNN-VAE, where its decoder network is trained only with one latent component alternatively and the remaining 24 components are set to zeros to examine the impact of each component for generating the patterns in EEG topo maps. The results show that each component contributes differently to capturing the generating aspects in topo maps. The empirical experiment was carried out using test data, with 10 samples chosen at random to assess the impact of each latent component in capturing the number of fundamental generative elements in spatially preserving EEG topographic maps. In order to analyse the results Using visual plausibility, ten images of test data and reconstructed images with active latent space components 0 are plotted, with findings clearly suggesting that each component is learning two to three patterns from those EEG topographic maps. These explanations can be used to gain the trust of stakeholders by demonstrating the visual plausibility of each latent component in capturing the generative components in EEG topographic maps.

5. Conclusion

Research on the interpretation of disentangled representations of VAE trained with spatially preserving EEG topographic maps is currently limited. A CNN-VAE decoder network is trained with alternatively one active latent component, and the remaining component is set to zero because the mean value is close to zero in the distribution learned from each latent component. The results with visual plausibility show that each component contributes differently to capturing and generating aspects in topo maps. Hence, this pipeline helps us understand each component of latent space responsible for activating a part of the brain region. Future studies will include understanding the decision of CNN-VAE through the interpretation of its latent space via clustering and visual plausibility, taking into account the signal-to-noise ratio and correlation values across the input and output of the architecture.

References

[1] C. Binnie, P. Prior, Electroencephalography., Journal of Neurology, Neurosurgery & Psychiatry 57 (1994) 1308–1319.

[2] E. W. Anderson, G. A. Preston, C. T. Silva, Using python for signal processing and visualization, Computing in science & engineering 12 (2010) 90–95.

[3] M. Taherisadr, M. Joneidi, N. Rahnavard, Eeg signal dimensionality reduction and classification using tensor decomposition and deep convolutional neural networks, in: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2019, pp. 1–6.

[4] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[5] S. M. Abdelfattah, G. M. Abdelrahman, M. Wang, Augmenting the size of eeg datasets using generative adversarial networks, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–6.

[6] J. F. Hwaidi, T. M. Chen, A noise removal approach from eeg recordings based on variational autoencoders, in: 2021 13th International Conference on Computer and Automation Engineering (ICCAE), IEEE, 2021, pp. 19–23.

[7] K. Li, J. Wang, S. Li, H. Yu, L. Zhu, J. Liu, L. Wu, Feature extraction and identification of alzheimer’s disease based on latent factor of multi-channel eeg, IEEE Transactions on Neural Systems and Rehabilitation Engineering 29 (2021) 1557–1567.

[8] X. Li, Z. Zhao, D. Song, Y. Zhang, J. Pan, L. Wu, J. Huo, C. Niu, D. Wang, Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks, Frontiers in neuroscience 14 (2020) 87.

[9] Z. Zheng, L. Sun, Disentangling latent space for vae by label relevant/irrelevant dimensions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12192–12201.

[10] T. Spinner, J. Körner, J. Görtler, O. Deussen, Towards an interpretable latent space: an intuitive comparison of autoencoders with variational autoencoders, in: IEEE VIS 2018, 2018.

[11] E. Mathieu, T. Rainforth, N. Siddharth, Y. W. Teh, Disentangling disentanglement in variational autoencoders, in: International Conference on Machine Learning, PMLR, 2019, pp. 4402–4412.

[12] N. Bryan-Kinns, B. Banar, C. Ford, C. Reed, Y. Zhang, S. Colton, J. Armitage, et al., Exploring xai for the arts: Explaining latent space in generative music (2022).

[13] A. Pati, A. Lerch, Attribute-based regularization of latent spaces for variational autoencoders, Neural Computing and Applications 33 (2021) 4429–4444.

[14] P. Cristovao, H. Nakada, Y. Tanimura, H. Asoh, Generating in-between images through learned latent space representation using variational autoencoders, IEEE Access 8 (2020) 149456–149467.

[15] F. Ding, Y. Yang, F. Luo, Clustering by directly disentangling latent space, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022, pp. 341–345.

[16] S. Mukherjee, H. Asnani, E. Lin, S. Kannan, Clustergan: Latent space clustering in generative adversarial networks, in: Proceedings of the AAAI conference on artificial

intelligence, volume 33, 2019, pp. 4610–4617.

[17] V. Prasad, D. Das, B. Bhowmick, Variational clustering: Leveraging variational autoencoders for image clustering, in: 2020 international joint conference on neural networks (IJCNN), IEEE, 2020, pp. 1–10.

[18] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE transactions on affective computing 3 (2011) 18–31.

[19] T. Ahmed, L. Longo, Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands, IEEE Access 10 (2022) 107575–107586. doi:10.1109/ACCESS.2022.3212777.

[20] A. V. Chikkankod, L. Longo, On the dimensionality and utility of convolutional autoencoder's latent space trained with topology-preserving spectral eeg head-maps, Machine Learning and Knowledge Extraction 4 (2022) 1042–1064.